Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 98 (2016) 419 – 424
Procedia Computer Science

International Workshop on Data Mining on IoT Systems (DaMIS16)

Mining geostatistics to quantify the spatial variability of certain soil flow properties

Gerardo SEVERINO[a,*], Maddalena SCARFATO[a], Gerardo TORALDO[b]

[a]*Department of* Agricultural Sciences, *University of Naples* - Federico II, *via Università* 100, *Portici* (NA), I-80055, Italy
[b]*Department of* Mathematics and Applications, *University of Naples* - Federico II, *via Cintia* (Monte S. Angelo), *Naples*, I-80126, Italy

Abstract

The functional dependence of the relative *unsaturated hydraulic conductivity* (UHC) $K_r(\psi) \equiv \exp(\alpha\psi)$ upon the matric potential ψ, [L], via the soil-dependent parameter α, [L^{-1}], has been traditionally regarded as a deterministic process (i.e. $\alpha \sim$ constant). However, in the practical applications one is concerned with flow domains of large extents where α undergoes to significant spatial variations as consequence of the disordered soil's structure. To account for such a variability (hereafter also termed as "heterogeneity") we adopt the mining geostatistical approach, which regards α as a *random space function* (RSF). To quantify the heterogeneity of α, estimates of local-values were obtained from $\sim$ 100 locations along a trench where an internal drainage test was conducted. The analysis of the statistical moments of α demonstrates (in line with the current literature on the matter) that the log-transform $\zeta \equiv \ln\alpha$ can be regarded as a structureless, normally distributed, RSF. An novel implementation of the present study in the context of the "Internet of Things" (IoT) is outlined.

Peer-review under responsibility of the Program Chairs

Keywords: soil; relative hydraulic conductivity; heterogeneity; mining geostatistics

1. Introduction

The challenging and very difficult task to develop modelling of flow and transport in soils of large extents has been undertaken only in the last decades by using a mining geostatistical approach[1,2]. The use of data-mining methods is due to the difficulties into quantifying the spatial distribution of the soil flow properties[3,4,5]. However, while a considerable effort has been invested to quantify the heterogeneity of certain soil properties, such as the Darcy's permeability coefficient[6], a very limited information about the spatial distribution of the α-parameter (relating the matric potential to the UHC) is available. Indeed, there have been only a limited number of studies[7,8,9,10] focusing on the spatial variability of α, and nevertheless they suffer from many limitations, the most important of which is about the extreme difficulty to carry out precise *in situ* measurements (somewhat similar to the analysis of water waves distribution[11]). In view of such shortcomings, the present paper aims at showing how to use a data-mining

* Corresponding author. Tel.: +39-081-2539426; fax: +39-081-2539412
E-mail address: gerardo.severino@unina.it

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.064

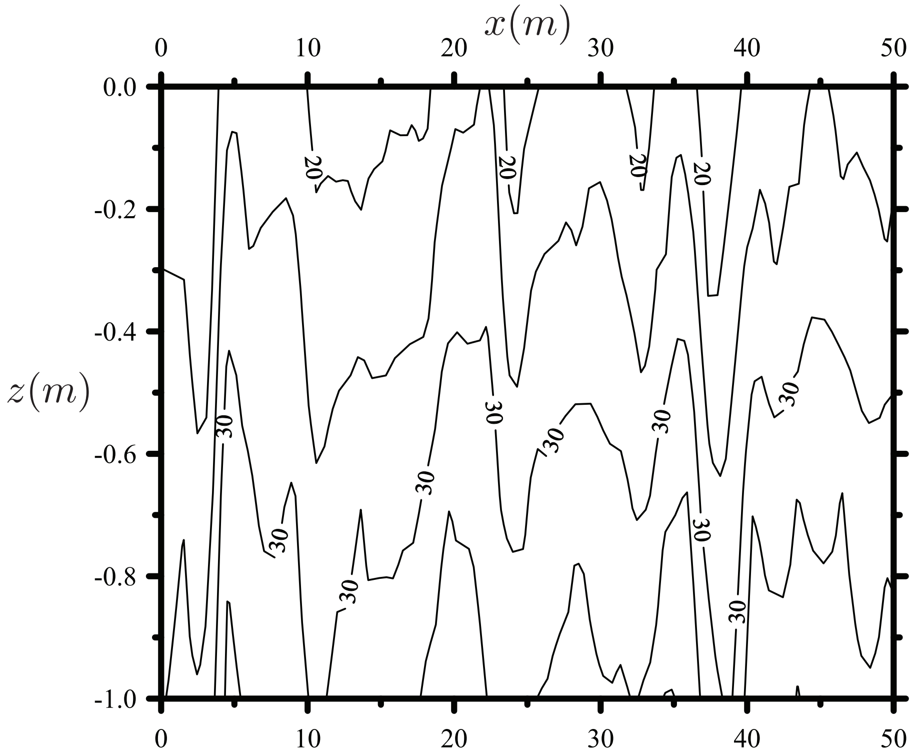


Figure 1. Distribution of the iso-values of λ_c (cm) along a vertical cross-section at the Ponticelli site (Naples, Italy); vertical exaggeration: 250/6.

(geostatistical) approach to quantify the spatial variability of the α-parameter. In addition, we believe that the present paper provides useful hints on how to combine devices/sensors and data in order to set up a compact web-tool (such as IoT) to gain quick analyses of complex (heterogeneous) environments, similarly to other studies concerning similar problems [12,11,13].

2. Characterization of the spatial variability of the α-parameter by means of the mining geostatistical approach: from theory to the practical use

The theoretical framework

The α-parameter is more than a curve-fitting number, since it is related to the soil's texture. Indeed, it has been demonstrated [8] that the characteristic length $\lambda_c \equiv \alpha^{-1}$, [L], is a measure of the importance of the capillary force relative to the gravitational one. More precisely, $\lambda_c \to 0$ implies that gravity dominates capillarity (coarse textured soils), and *viceversa* (fine textured soils). Since, the soil's texture is highly variable from point to point in the soil, a tantamount degree of variability is detected into the values taken by the α-parameter. This is clearly seen in the Figure 1 that shows the contour levels of λ_c (cm) along a vertical cross-section in a trench.

A detailed characterization of the spatial distribution of α (and more generally of any soil flow property) via the so-called "standard approach" (i.e. by collecting samples in the field and subsequently determining local values) requires: i) considerable time, and ii) great expense/effort, therefore rendering such an avenue practically impossible. A viable (and widely accepted) alternative is to treat α as a "stochastic process in the space" or equivalently a RSF [14,6]. As a consequence, characterization of the heterogeneity of α is cast within the more general approach of the data mining methods.

Thus, the value of α at any $\boldsymbol{x}$ is regarded as one out-coming related to the many possible geologic materials that might have been generated there. As a consequence, $\alpha \equiv \alpha(\boldsymbol{x};\Omega)$ becomes a random variable. The symbol Ω refers to the sample space, which is generally dependent upon the position $\boldsymbol{x}$. Likewise, if α is measured at different positions $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ then the values $\alpha_i \equiv \alpha_i(\boldsymbol{x}_i;\Omega_i)$ $(i = 1, \ldots, k)$ are random variables, each one characterized by a (generally position dependent) *probability density function* (PDF). In addition, the possibility of finding any sequence of α-values at a certain $\boldsymbol{x}$ depends not only upon the PDF itself, but also on those PDFS at other positions. In the context of the mining geostatistics, the probability of finding such a sequence is given by the *joint probability density function*. Thus, any sequence of α-values at different points is viewed as a possible out-coming of the sample space of a joint PDF, and it is usually termed as *single realization*. As a matter of fact, determining the occurrence of any realization requires the knowledge of the joint PDF. Unfortunately, this latter is not an accessible information since in the practice only a single realization (the one obtained by the sampling) is available, and therefore one must resort to some simplifying assumptions, i.e.*stationarity* and *ergodicity*. Stationarity implies that the joint PDF is translationally invariant, whereas ergodicity enables one to infer the joint PDF by means of a single realization[6]. The pragmatic approach adopted in *Hydrology*, and in line with the statistical continuos theories, is to derive moments of interest for the flow variables and to check the applicability of these two assumptions only *ex post*. In terms of moments, stationarity requires the space invariance of "all" the moments: a very stringent assumption. Since, in the practical applications one is mainly interested into the first and second order moments of the flow/transport processes[15,16,17,18,19,20,21], the stationarity of the input variables is replaced by the stationarity up to the second order (*weak stationarity*). Thus, the pair "mean and covariance" becomes the tool to characterize the spatial variability of α. Nevertheless, it is important to emphasize that the knowledge of the mean and covariance does not specify the α-values at any $\boldsymbol{x}$, but it rather provides a way to quantify how widely the α-values spread around the mean, and how these values are spatially correlated.

Results and discussion

In the present paper local measurements of α were obtained by means of a field-scale drainage experiment[22] at the Ponticelli site (Naples, ITALY). Along a transect (50m long) 40 verticals (1.25 m apart) were chosen, and for each of them the pair (ψ, K_r) was measured at three depths (z = 30, 60, 90 cm). Hence, from the 40×3 available pairs, the α-parameter was obtained via a best fitting procedure[23], and the resulting spatial distribution is shown in the Figure 1. The cumulative distribution function of $\alpha\,(\mathrm{m}^{-1})$ (red) together with its logarithmic transform $\zeta \equiv \ln\alpha$ (black) is depicted in the Figure 2. At a first glance, it is seen that the empirical distribution (discrete symbols) exhibits a larger deviation from the normal distribution, whereas deviations from the log-normal one are smaller. This is quantitatively confirmed by inspection of Table 1 where we summarize (among the other) the result of the hypothesis (Kolmogorov-Smirnov) normality test.

statistics	$\alpha^\dagger$	$\zeta \equiv \ln\alpha$
mean	3.88	1.33
variance	1.01	$5.80 \cdot 10^{-2}$
D	1.03	$5.97 \cdot 10^{-1}$

†values of α are in m^{-1}

Table 1. Estimates of the: i) mean, and ii) variance, together with the test (D) of normal/log-normal (null) hypothesis.

The problem of quantifying the spatial structure (i.e.the covariance, in the present study) of α is rather complicated, even when measurements are numerous. The identification process should involve several steps: i) an hypothesis about the functional model of the covariance, ii) estimates of the parameters of such a model, and iii) a model validation test. However, the problem of selecting the most appropriate model remains to some extent in the realm of the practical applications[6]. The prevailing approach is the pragmatic one: select a model for its practicality/versatility as well as its performance in similar situations. Nevertheless, it is important in view of the subsequent analysis to discuss some general properties of the covariance function $C \equiv C(\boldsymbol{x})$. Thus, the value $C(0)$ is the so-called "structured variance", and it provides information about the spread of the α-values around the mean. For $|\boldsymbol{x}| \neq 0$, the value $C(\boldsymbol{x})$ is a measure of the correlation between the α-values at two points separated by the distance $|\boldsymbol{x}|$. More precisely, the higher is $|\boldsymbol{x}|$ the smaller the correlation. Of particular interest is the concept of *integral scale*, $\mathcal{I}_\alpha$. Roughly speaking the integral scale, [L], represents the distance over which two values of α cease to be correlated[14]. A frequently encountered case

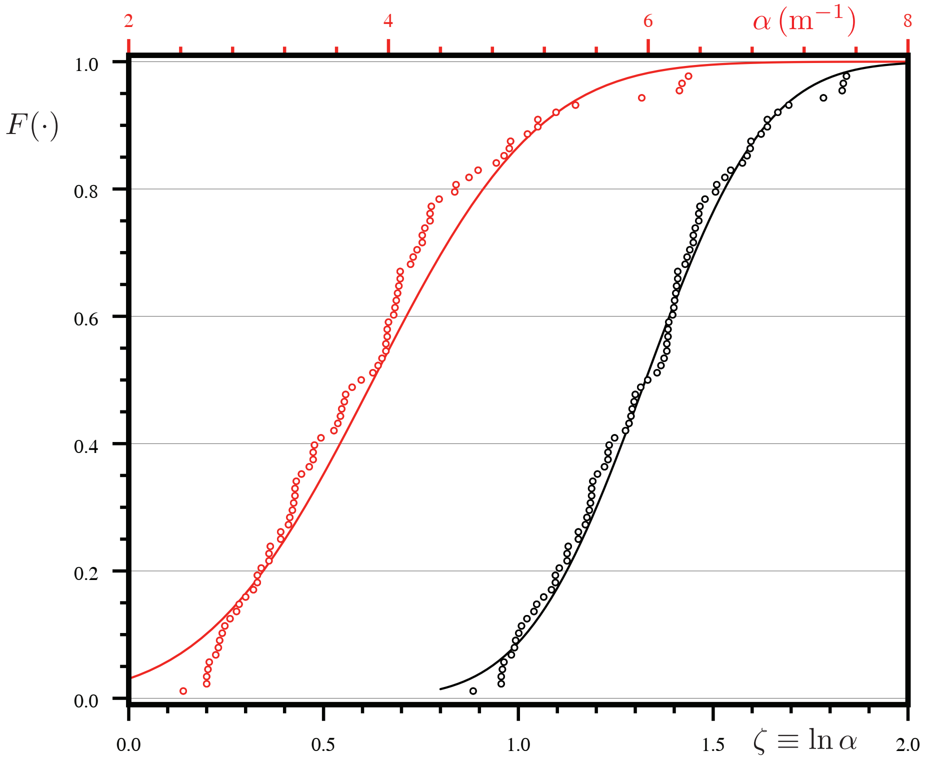


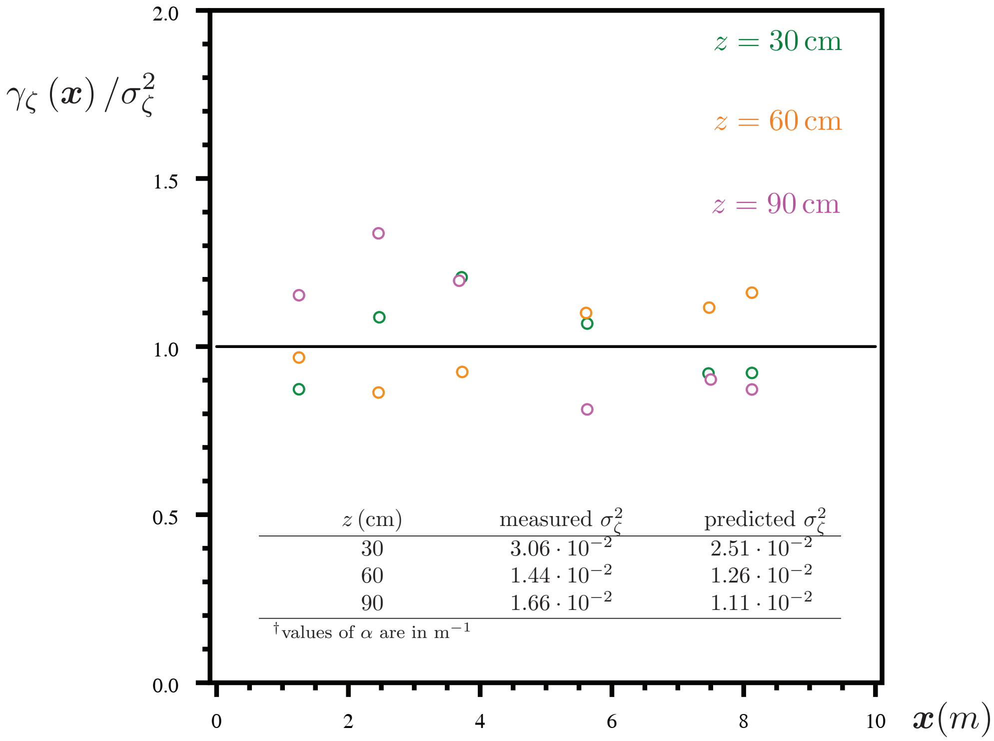
Figure 2. Cumulative distribution function of $\alpha\,(\mathrm{m}^{-1})$ (red), and its log-transform $\zeta \equiv \ln \alpha$ (black). Discrete symbols and continuous lines refer to the empirical distribution and to the models, respectively.

is that of zero intregral scale, i.e. $\mathcal{I}_\alpha \to 0$. In this case the geological formation is characterized by a complete lack of spatial correlation, i.e. $C(\boldsymbol{x}) \simeq 0$ for any $|\boldsymbol{x}| \neq 0$, and this is known as *stochastic structureless process*. In such a circumstance, it is convenient to deal with the variogram $\gamma \equiv \gamma(\boldsymbol{x})$ [14,6]. Generally, the variogram γ (whose computation is straightforward) is of wider applicability as compared with the covariance, since its applicability does not require the stationarity hypothesis in a strict sense. Nevertheless, for a stationary process one can easily demonstrate [14,6] that $\gamma(\boldsymbol{x}) \equiv C(0) - C(\boldsymbol{x})$. As a consequence, for a stochastic, stationary, structureless process the variogram in practice coincides with the structured variance, i.e. $\gamma(\boldsymbol{x}) \simeq C(0)$. Thus, the use of the variogram is very useful to visualize whether any stochastic process is structureless. The experimental scaled-variograms γ/σ_ζ^2 at three different depths found for the transect in Figure 1 is plotted in the Figure 3. The fact that $\gamma/\sigma_\zeta^2 \sim 1$ supports the assumption of a spatial lack of correlation, and concurrently for the geological formation at stake the RSF α can be regarded as a structureless stochastic process.

3. Concluding remarks and highlights toward an implementation in the context of the IoT

A preliminary analysis of a field scale drainage test suggests that the log-transform $\zeta \equiv \ln \alpha$ of the parameter appearing into the relative UHC: $K_r \equiv \exp(\alpha\psi)$, characterizing the length of the capillary force acting into unsaturated porous media (soils), can be modeled as a stationary RSF of zero integral scale (i.e. a structureless stochastic process). The most important consequence in view of the applications is that the covariance of ζ can be approximated by a white noise signal in the horizontal plane.

Before concluding, we wish to highlight here an application of the presented material which can be easily implemented



z (cm)	measured σ_ζ^2	predicted σ_ζ^2
30	$3.06 \cdot 10^{-2}$	$2.51 \cdot 10^{-2}$
60	$1.44 \cdot 10^{-2}$	$1.26 \cdot 10^{-2}$
90	$1.66 \cdot 10^{-2}$	$1.11 \cdot 10^{-2}$

†values of α are in m^{-1}

Figure 3. Scaled variogram $\gamma_\zeta/\sigma_\zeta^2$ at the three measuring depths along the horizontal distance $\boldsymbol{x}$. The measured *vs* predicted structured variances σ_ζ^2 are shown in the insert.

in the IoT-context. Indeed, data-driven agricultural technologies are rapidly becoming a tool of large use, and in particular they allow one to design a site-specific management plan (precision-agriculture). In particular, a majority of precision-agriculture strategies rely on statistical analyses (or image processing) of indirect measurements of soil conditions obtained, for example, by satellites, unmanned aircraft or other means of remote sensing. Various (above-ground) parameters related to crop conditions can be effectively monitored with wireless sensor networks. Thus, the utility of our approach comes from the use of dynamic real-time forecasting of the quantity and quality of soil water to guide the field irrigation. This forecasting will be facilitated and informed by *in situ* measurements of water content obtained with spatially distributed autonomous and automated sensors along an IoT-approach [24].

Acknowledgements

The Authors acknowledge the support from the CAISIAL Department within the MODige-project (E72C14000170005). Gerardo Severino acknowledges the kind assistance of his students (class: *Protection of environmental water resources*) during the data-processing.

References

1. Comegna, A., Coppola, A., Comegna, V., Severino, G., Sommella, A., Vitale, C.. State-space approach to evaluate spatial variability of field measured soil water status along a line transect in a volcanic-vesuvian soil. Hydrology and Earth System Sciences 2010;14(12):2455–2463.
2. Severino, G., Comegna, A., Coppola, A., Sommella, A., Santini, A.. Stochastic analysis of a field-scale unsaturated transport experiment. Advances in Water Resources 2010;33(10):1188–1198.
3. Severino, G., Monetti, V., Santini, A., Toraldo, G.. Unsaturated transport with linear kinetic sorption under unsteady vertical flow. Transport in porous media 2006;63(1):147–174.
4. Severino, G., Santini, A., Monetti, V.M.. Modelling water flow water flow and solute transport in heterogeneous unsaturated porous media. In: Advances in Modeling Agricultural Systems. Springer; 2009, p. 361–383.
5. Severino, G., Coppola, A.. A note on the apparent conductivity of stratified porous media in unsaturated steady flow above a water table. Transport in porous media 2012;91(2):733–740.
6. Rubin, Y.. Applied Sthocastic Hydrogeology. Oxford: Oxford University Press; 2003.
7. Russo, D., Bouton, M.. Statistical analysis of spatial variability in unsaturated flow parameters. Water Resources Research 1992;28(7):1911–1925.
8. White, I., Sully, M.. On the variability and use of the hydraulic conductivity alpha parameter in stochastic treatments of unsaturated flow. Water resources research 1992;28(1):209–213.
9. Ragab, R., Cooper, J.. Variability of unsaturated zone water transport parameters: implications for hydrological modelling. 1. in situ measurements. Journal of hydrology 1993;148(1-4):109–131.
10. Russo, D., Russo, I., Laufer, A.. On the spatial variability of parameters of the unsaturated hydraulic conductivity. Water Resources Research 1997;33(5):947–956.
11. Farina, R., Dobricic, S., Storto, A., Masina, S., Cuomo, S.. A revised scheme to compute horizontal covariances in an oceanographic 3d-var assimilation system. Journal of Computational Physics 2015;284:631–647.
12. Cuomo, S., De Michele, P., Galletti, A., Piccialli, F.. A cultural heritage case study of visitor experiences shared on a social network. In: 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC). IEEE; 2015, p. 539–544.
13. Cuomo, S., De Pietro, G., Farina, R., Galletti, A., Sannino, G.. A revised scheme for real time ecg signal denoising based on recursive filtering. Biomedical Signal Processing and Control 2016;27:134–144.
14. Dagan, G.. Flow and Transport in Porous Formation. New York: Springer-Verlag; 1989.
15. Severino, G., Santini, A., Sommella, A.. Steady flows driven by sources of random strength in heterogeneous aquifers with application to partially penetrating wells. Stochastic Environmental Research and Risk Assessment 2008;22(4):567–582.
16. Fiori, A., Boso, F., de Barros, F.P., De Bartolo, S., Frampton, A., Severino, G., et al. An indirect assessment on the impact of connectivity of conductivity classes upon longitudinal asymptotic macrodispersivity. Water resources research 2010;46(8).
17. Severino, G.. Stochastic analysis of well-type flows in randomly heterogeneous porous formations. Water Resources Research 2011;47(3).
18. Severino, G., Santini, A., Sommella, A.. Macrodispersion by diverging radial flows in randomly heterogeneous porous media. Journal of contaminant hydrology 2011;123(1):40–49.
19. Severino, G.. Macrodispersion by point-like source flows in randomly heterogeneous porous media. Transport in porous media 2011;89(1):121–134.
20. Severino, G., De Bartolo, S., Toraldo, G., Srinivasan, G., Viswanathan, H.. Travel time approach to kinetically sorbing solute by diverging radial flows through heterogeneous porous formations. Water Resources Research 2012;48(12).
21. Severino, G., De Bartolo, S.. Stochastic analysis of steady seepage underneath a water-retaining wall through highly anisotropic porous media. Journal of Fluid Mechanics 2015;778:253–272.
22. Severino, G., Santini, A., Sommella, A.. Determining the soil hydraulic conductivity by means of a field scale internal drainage. Journal of Hydrology 2003;273(1):234–248.
23. Gómez, S., Severino, G., Randazzo, L., Toraldo, G., Otero, J.. Identification of the hydraulic conductivity using a global optimization method. agricultural water management 2009;96(3):504–510.
24. Chianese, A., Marulli, F., Piccialli, F., Benedusi, P., Jung, J.E.. An associative engines based approach supporting collaborative analytics in the internet of cultural things. Future Generation Computer Systems 2016;.